# Guangwai Lancaster Chinese Learner Corpus



Guangwai Lancaster Chinese Learner Corpus of L2 Chinese production is a result of collaboration between Guangdong University of Foreign Studies (GDUFS) and Lancaster University. The corpus contains over one million running words of both written and spoken language. It is a balanced sample that covers three proficiency levels: beginner, intermediate and advanced, thus providing a unique insight into the lexical and grammatical development of L2 Chinese.

#### Key features of the corpus include:

- 1.2 million words
- 886 L2 Chinese speakers from 80 countries
- Three proficiency levels
- Fully error tagged
- XML format with rich metadata



## Who are the speakers in the corpus?

The L2 Chinese learners in the corpus are students of Chinese as a second/foreign language at GDUFS, China. They come from a large variety of countries worldwide. L1 Chinese speakers in the corpus are staff and students at GDUFS.



Photos by Guangdong University of Foreign Studie

# What spoken interactions and written texts are included?

The data (spoken and written) was collected from exams and tutorial sessions carried out at GDUFS. The written texts are usually short pieces and essays on a given topic. Spoken data comprise interactions typically between native and non-native speakers of Chinese and involve one, two or multiple speakers.

### What insights can the corpus give us about L2 Chinese?







